会計ドメインにおける質問応答のための LLM を用いた解説ページ順位付け

飯田頌平[1] 古俣槙山[2] 三田寺聖[2] 長谷川遼[2]
宇津呂武仁[2] 林友超[1] 宍戸里絵[1]
[1]弥生株式会社 次世代本部 R&D 室
[2]筑波大学大学院 システム情報工学研究群 知能機能システム学位プログラム

概要

大規模言語モデル（LLM）は幅広い分野で応用が進んでおり，会計ドメインにおいても活用が期待されている．そこで本論文では，弥生株式会社の持つ会計業務に関する質問の解説ページを収集し，検索システム構築および精度評価に用いることのできる弥生 QA v1 データセットを構築した．このデータセットを用いて既存のベクトル検索手法の評価を行い，さらにその検索結果を LLM によって順位付けした．この結果，順位付けを実施しない従来手法に比べて正答率の向上が見られ，会計ドメインにおいて有効となる検索手法を示した．

1 はじめに

会計とは資金の流れを記録し報告することで，事業の継続にあたり必要な業務である．しかし中小法人や個人事業主では専門の人員を配置していないケースや事業者の会計知識が不足しているケースがあり，その結果，適切な税務手続きに失敗する例やキャッシュの枯渇を招く例が見られ，時には事業の継続が不可能となる事例もある．

そのような社会的課題に取り組むため，弥生株式会社では会計をはじめとした様々なバックオフィス業務に関連するソフトウェアを展開するほか，専門家によって体系化された会計知識のナレッジを公開することで，あらゆるスモールビジネスの事業者の会計への意識を啓発している．しかし，専門知識を持たない一般人が会計業務について調べることは難しく，どのように情報を検索すればよいかわからないという声も挙がっている．

このような背景から，近年大きく発展を遂げている大規模言語モデル (LLM) [18] を用い，自然言語による質問から会計ドメインの専門知識を検索するシステムが期待される．そこで本論文では，会計ドメインに特化したデータセットを構築して既存のベクトル検索手法における性能を評価し，さらに検索結果を LLM により順位付けする新手法を提案した．本論文の貢献は主に以下の要素から構成される．

弥生 QA v1 データセット 弥生株式会社の持つ会計業務に関するナレッジから構築された，質問とその回答の根拠となる解説ページの対から構成されるコーパス．

LLM を用いた順位付け FAISS[1)] によるナレッジの検索結果を元に LLM に順位付けを行う指示を与えることで，より質問に対して適切な検索結果を得られる手法．

弥生 QA v1 データセットを用い Top-k Accuracy ($k = 1$) において LLM による質問応答手法の定量的評価を実施したところ，FAISS による従来の検索手法では 74.23 ポイントであった一方，提案手法では 86.15 ポイントに改善し，会計という専門性の高いドメインにおいても高い正答率を示した．

2 関連研究

文献 [4] では，会計および財務ドメインに関する LLM の研究事例について包括的な調査を行った．それによると，文献 [1, 3, 17] では税務や監査における LLM の能力に関する調査が実施された．さらに，文献 [2] では LLM を Robotic Process Automation (RPA) に組み合わせることで，会計の生産性と効率性を高めることを提案している．また BloombergGPT [15] や FinGPT [16] といった金融に関連したコーパスを用いた学習によりドメイン特化型の LLM を構築する試みが実施されている．そのほか，文献 [6] は金融ドメインにおける LLM の性能評価のためのベンチマークを構築し，文献 [9] では株価の推移を示す

1) https://github.com/facebookresearch/faiss

用語を LLM によって選択する手法が提案された．

一方，Retrieval-Augmented Generation (RAG) [8] のように情報源を参照することでより適切な応答を可能とする手法も注目されている [5]．LLM の抱える問題として正確ではない回答を生成する「幻覚」[7] が挙げられるが，RAG によって「幻覚」を抑制することが可能だと知られている [11]．特に会計・財務ドメインにおいては専門知識の複雑さや法令のアップデートといった「幻覚」が発生しやすい背景があり，また誤った情報をもとに会計業務を実施することで経済的な損失や法的リスクを招く恐れがあるため，一般的なドメインよりもさらに注意深く「幻覚」に対処する必要がある [12]．このような問題に対する直近の取り組みとして，RAG によって抽出した金融知識を反映させる研究 [10] などが挙げられる．

3　弥生 QA v1 データセット

弥生 QA v1 データセットは，Web 上で公開されている会計情報の解説ページと，その記事を検索するための質問から構成される．以下，解説ページおよび質問の詳細について述べる．

3.1　弥生株式会社の提供する解説ページ

弥生株式会社では，様々な媒体上で会計や金融，経営に関する知識の啓発を実施している．その中でも特に下記のコンテンツを対象とし，弥生 QA v1 データセットに用いることとした．

お役立ち情報　会計ドメインの基礎知識を体系的に解説したもの．経理お役立ち情報[2] をはじめ，確定申告や給与計算など，11 カテゴリから構成される．詳細は表 1 を参照．

資金調達ナビ　弥生が展開する「資金調達手段を探せる・学べる・専門家を頼れる」サービス[3]．資金調達に必要となる知識を専門家が解説したものが掲載されており，銀行からの融資や行政からの助成金・補助金といった，ファイナンスに関する様々な情報を掲載している．

弥報　経営者向けに幅広い情報を発信するビジネス情報メディア[4]．主要なコンテンツは「顧客獲得・売上アップ」「人材（採用・育成・定着）」「事業成長・経営力アップ」「経営ノウハウ＆トレンド」の 4 トピックから構成される．

2）https://www.yayoi-kk.co.jp/kaikei/oyakudachi/
3）https://shikin.yayoi-kk.co.jp/study/index.html
4）https://media.yayoi-kk.co.jp/

以上の Web ページを収集し，合計 2,454 件の記事をデータセットとして使用した（表 1）．なお，記事の件数は 2025 年 1 月 6 日時点のものである．

表 1　弥生 QA v1 データセット

種別	記事数	質問数
経理お役立ち情報	295	260
確定申告お役立ち情報	258	-
青色申告お役立ち情報	46	-
請求書作成お役立ち情報	196	-
給与計算お役立ち情報	208	-
副業お役立ち情報	60	-
起業・開業お役立ち情報	220	-
税理士相談お役立ち情報	20	-
M&A・事業承継お役立ち情報	46	-
電子帳簿保存法お役立ち情報	55	-
インボイス制度お役立ち情報	107	-
資金調達ナビ	220	190
弥報	723	594
合計	2,454	1,044

3.2　LLM を用いた質問の生成

まず「経理お役立ち情報」の 295 記事から，人手評価により質問の生成に使用可能な 260 記事を抽出した．使用不可能とした 35 記事は，次のいずれかの理由に該当するものである．

- 複数の要素をまとめたもので，質問内容への言及が記事に占める割合が低い記事．
- 類似する内容について言及した記事が複数存在し，検索時の混同が極めて起こりやすいと考えられる記事．
- 動画への誘導などを目的としており，Web ページ内容の情報量が不十分である記事．

次に抽出された各記事に対し，回答を得られるような質問を生成するよう指示を与え，LLM によって質問を生成した．LLM としては OpenAI API の GPT-4o [5] モデルを用いた．なお指示として与えたプロンプトでは，次のような要件を定めた．

- 3 通りの質問候補を出力する．
- 一般人による質問を想定する．
- 専門用語の使用を可能な限り避ける．

これにより得られた 3 通りの質問候補の中から，最も適した質問を選択した．記事の選定時と同様に

5）https://platform.openai.com/docs/models#gpt-4o

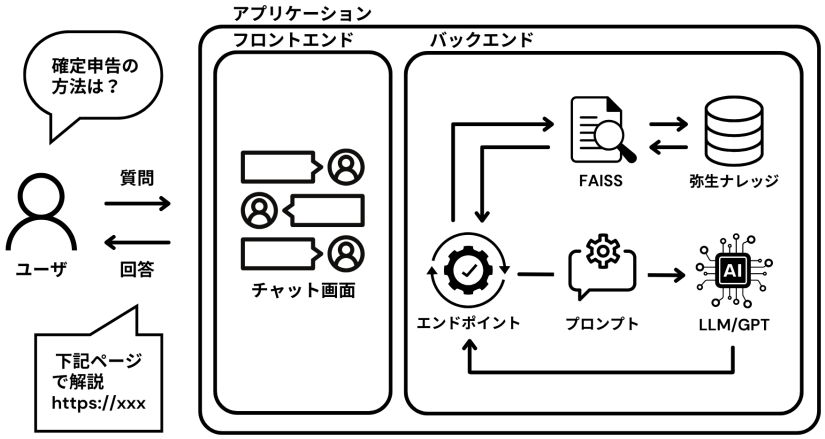

図 1　質問応答の枠組みの全体像

評価者による判断に基づき，また質問選択の過程において，次の特徴を持つ質問は不採用とした．

- 記事との対応が取れない質問．
- 参照不可能なコンテキストを含む質問．
- 会計ソフトの使い方に関する質問．

会計ソフトの使い方に関するものは，多くの記事で同様の内容が含まれるため除外することとした．さらに，生成された 3 質問がすべて不適切であったものに関しては，記事の内容自体は使用可能であっても使用不可能な記事として扱った．

上記の質問生成プロセスを「経理お役立ち情報」の他，「資金調達ナビ」「弥報」に対しても実施し，合計で表 1 に示す記事と質問の対を得た．この対を順位付け手法の評価用データとして用いる．

4　質問応答の枠組み

ユーザが質問を実施し本論文の手法に基づいた回答を得るまでの枠組みを図 1 に示す．まずユーザがチャット画面に入力した質問を元に，FAISS によって回答の根拠となるページを検索する．その後，LLM を用いてページの順位付けを行い，回答をチャット画面へと返すことで，質問から回答までの一連の振る舞いを実現する．

4.1　FAISS を用いた弥生 QA v1 データの検索

埋め込みにより類似度検索を実施するライブラリ FAISS により，質問文との類似度の高い解説ページの順位付けを行う．埋め込みとしては ruri-large [13] を利用した．さらに埋め込みの類似度を測るための指標として，コサイン類似度を使用した．

4.2　LLM を用いた順位付け

本論文では，FAISS によって抽出された検索結果に対し，LLM を用いて再度順位付けを行う手法を提案する．LLM としては，Google 社の提供する gemini-1.5-pro [6] を利用する．LLM にはプロンプトを用いて質問文との類似度の高い解説ページの順位付けを行うタスクを実施させる．プロンプトには次の情報を与える．

指示　質問と記事リストを入力として，回答例を参照しつつ質問の回答が記載されたページを答えるよう指示を記載したもの．

回答例　JSON フォーマットに従い順位，記事 ID，記事タイトルをキーとして指定したもの．

質問　3.2 節で生成した質問．

記事リスト　4.1 節で抽出した FAISS による検索結果上位 5 件の記事 ID，記事タイトル，本文．

さらに，Chain-of-Thought (CoT) [14] の概念を取り入れ，順位付けの理由を生成するよう指示を追加する手法も実施した．以下，5 章では，LLM による順位付け手法とさらに CoT を取り入れた手法について，それぞれ評価実験を行った結果を記載する．

5　評価実験

4 章で考案した枠組みを評価するため，3.2 節で生成した質問から 3.1 節で収集した記事を検索するタスクを実施した．表 2 に各評価用データセット毎の実験結果を示す．評価指標として Top-k Accuracy を使用した．ベースラインである ruri-large を用いた

6）https://deepmind.google/technologies/gemini/pro

表 2　評価結果

評価データ	手法	Top-k Accuracy[%]				
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
経理お役立ち情報	FAISS (ruri-large)	74.23	86.54	93.46	95.38	**96.15**
	LLM を用いた順位付け (gemini-1.5-pro)	81.54	93.85	**96.15**	**96.15**	**96.15**
	LLM を用いた順位付け + CoT (gemini-1.5-pro)	**86.15**	**94.61**	95.77	**96.15**	**96.15**
資金調達ナビ	FAISS (ruri-large)	60.53	80.00	88.42	93.68	94.73
	LLM を用いた順位付け (gemini-1.5-pro)	69.47	**82.63**	**91.58**	93.16	**94.74**
	LLM を用いた順位付け + CoT (gemini-1.5-pro)	**72.63**	**82.63**	89.47	**94.21**	**94.74**
弥報	FAISS (ruri-large)	72.39	84.34	89.56	91.92	**93.43**
	LLM を用いた順位付け (gemini-1.5-pro)	72.22	87.04	**90.24**	**92.92**	**93.43**
	LLM を用いた順位付け + CoT (gemini-1.5-pro)	**78.62**	**87.37**	**90.24**	92.42	**93.43**

FAISS による検索結果と比較すると，LLM による順位付け手法は概ね高い正答率を示した．また CoT を用いることでさらに正答率が改善し，特に $k=1$ のとき，いずれの評価用データセットを用いた実験においても他の手法に比べ高い正答率を示した．

6　分析

6.1　順位付けによって改善したページ

「経理お役立ち情報」の評価データにおける事例を付録の表 4 に示す．1 例目では「レジのお金が計算と合わない」という質問に対し，「ロス」という文言へ適切に結びつけることによって，LLM を用いた順位付け手法と CoT を加えた手法では，質問の回答としてより相応しい記事を検索できた．2 例目においては，LLM を用いるだけではベースラインを悪化させてしまったところ，CoT により理由を考慮させることで正解を得られた．さらに 3 例目においては，CoT を取り入れた場合にのみ，質問で言及された「定額法と定率法」から背景の「減価償却」を紐付け正解を得られた結果となった．

次に「資金調達ナビ」の評価データにおける事例を付録の表 5 に示す．「資金調達ナビ」の記事はファイナンスの分野に特化しており，類似した記事が多くなる傾向が見てとれた．そのため評価結果（表 2）においても $k=1$ と $k=5$ の差分が他のデータよりも大きく，類似記事の中から正解を選ぶ難易度の高さが示されている．

最後に「弥報」の評価データにおける事例を付録の表 6 に示す．「弥報」においては様々なトピックのデータが含まれるため，推論結果の分散が大きくなる傾向があった．評価結果（表 2）においても $k=1$ の際に LLM を用いた順位付け手法が FAISS による検索手法よりも低い正答率を示している．しかし CoT を組み合わせ，順位付けの理由を考慮させることによって，このようなデータにおいても安定して高い正答率を得られることが示された．

6.2　逆検索

4 節の手法は検索結果を元に順位付けを行うものであるが，検索結果から LLM によって回答を直接生成し，その回答に類似した記事を FAISS で再検索する手法を実施した．

表 3　評価結果: 逆検索（経理お役立ち情報）

手法	Top-k Accuracy[%]				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
逆検索	71.53	89.23	92.31	94.61	**96.54**

「経理お役立ち情報」の評価データに対し実施した逆検索手法の評価結果（表 3）を確認すると，$1 \leq k \leq 4$ においては順位付けを実施する提案手法（表 2）の方が，$k=5$ の場合には逆検索の方が高い正答率を示した．一般的に逆検索手法は正答率が低くなるものの，FAISS による検索結果の上位 5 件に正しい回答が含まれていない場合においても正解を検索し得るという利点が示された．

7　おわりに

本論文では，会計ドメインにおける質問応答に有用なデータセットを構築し，さらに応答の根拠となる解説ページの順位付けにおける新たな手法を提案した．この結果，LLM による順位付け手法は既存の手法より優れた値を示し，会計ドメインの情報検索において効果的であることを示した．

謝辞

本論文は筑波大学および弥生株式会社の共同研究の一環として執筆いたしました．本論文の執筆にあたりご協力いただきました，筑波大学および弥生株式会社の関係者の皆様に深くお礼申し上げます．

参考文献

[1] B. Alarie, K. Condon, S. Massey, and C. Yan. The rise of generative AI for tax research. **Tax Notes Federal**, Vol. 179, pp. 1509–1522, 2023.

[2] D. Beerbaum. Generative artificial intelligence (GAI) ethics taxonomy-applying chat GPT for robotic process automation (GAI-RPA) as business case. **Available at SSRN 4385025**, pp. 1–20, 2023.

[3] G. Y. Choi and A. Kim. Economic footprints of tax audits: A generative AI-driven approach. **Chicago Booth Research Paper**, pp. 1–83, 2023.

[4] M. M. Dong, T. C. Stratopoulos, and V. X. Wang. A scoping review of ChatGPT research in accounting and finance. **International Journal of Accounting Information Systems**, Vol. 55, pp. 100715–100744, 2024.

[5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang. Retrieval-augmented generation for large language models: A survey. **arXiv e-print**, pp. 1–21, 2023.

[6] 平野正徳. 金融分野における言語モデル性能評価のための日本語金融ベンチマーク構築. 言語処理学会第 30 回年次大会発表論文集, pp. 1570–1574, 2024.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. **ACM Computing Surveys**, Vol. 55, No. 12, pp. 1–38, 2023.

[8] P. Lewis, E. Perez, P. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. **Proc. 34th NIPS**, Vol. 33, pp. 9459–9474, 2020.

[9] 西田隼輔, 宇津呂武仁. 株価変動に対する大規模言語モデルを用いた株式用語選択. 言語処理学会第 30 回年次大会発表論文集, pp. 1559–1564, 2024.

[10] 西脇一尊, 大沼俊輔, 工藤剛, 門脇一真. ファイナンシャル・プランニングの自動化に向けた GPT-4 及び RAG の性能評価. 言語処理学会第 30 回年次大会発表論文集, pp. 1575–1580, 2024.

[11] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston. Retrieval augmentation reduces hallucination in conversation. In **Findings of EMNLP**, pp. 3784–3803, 2021.

[12] S. M. T. I. Tonmoy, S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das. A comprehensive survey of hallucination mitigation techniques in large language models. **arXiv e-prints**, pp. 1–19, 2024.

[13] H. Tsukagoshi and R. Sasano. Ruri: Japanese general text embeddings. **arXiv e-prints**, pp. 1–14, 2024.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Proc 36th NIPS**, Vol. 35, pp. 24824–24837, 2022.

[15] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann. BloombergGPT: A large language model for finance. **arXiv e-prints**, pp. 1–76, 2023.

[16] H. Yang, X. Y. Liu, and C. D. Wang. FinGPT: Open-source financial large language models. **arXiv e-prints**, pp. 1–7, 2023.

[17] L. Zhang. Four tax questions for ChatGPT and other language models. **Tax Notes Federal**, Vol. 179, pp. 969–974, 2023.

[18] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Y. Nie, and J. R. Wen. A survey of large language models. **arXiv e-prints**, pp. 1–97, 2023.

A　評価結果の具体例

表 4 に「経理お役立ち情報」の評価データを用いた実験結果から, 質問とそれに対する順位が最上位であった記事の具体例を記載する.

表 4　「経理お役立ち情報」の評価データにおける Top-1 の具体例

質問	手法	記事タイトル	分類
レジのお金が計算と合わない場合, どのように経理処理を行うべきですか？	FAISS	飲食店における経理業務で重要なポイント...	×
	LLM	【飲食業・小売業】ロスがあった場合の経理処理	○
	CoT	【飲食業・小売業】ロスがあった場合の経理処理	○
不動産賃貸業における経理業務で注意すべき点は何ですか？	FAISS	不動産業における経理業務とは？...	○
	LLM	決算月にやっておきたい 5 つの作業	×
	CoT	不動産業における経理業務とは？...	○
定額法と定率法の違いは何ですか？...	FAISS	定額減税による所得税の減税額は 3 万円！...	×
	LLM	固定資産台帳とは？決算書作成に必要な...	×
	CoT	減価償却の計算方法とは？定額法・定率法...	○

表 5 に「資金調達ナビ」の評価データを用いた実験結果から, 質問とそれに対する順位が最上位であった記事の具体例を記載する.

表 5　「資金調達ナビ」の評価データにおける Top-1 の具体例

質問	手法	記事タイトル	分類
日本政策金融公庫の創業融資を受けるために, どのような追加資料を用意することが推奨されていますか？	FAISS	...「創業融資の際に必要な資料」	×
	LLM	...「創業融資申請書類（日本公庫）の書き方」	○
	CoT	...「創業融資申請書類（日本公庫）の書き方」	○
有限会社と合同会社の違いは何ですか？特に増資に関する手続きについて教えてください.	FAISS	株式会社以外の増資	○
	LLM	有限会社と株式会社の違いは？...	×
	CoT	株式会社以外の増資	○
補助金や助成金を受け取った後, どのような報告や手続きが必要になりますか？	FAISS	新規事業立ち上げに使える補助金・助成金 9 選	×
	LLM	補助金と助成金の違いは？...	×
	CoT	補助金・助成金の申請から受給までの流れ	○

表 6 に「弥報」の評価データを用いた実験結果から, 質問とそれに対する順位が最上位であった記事の具体例を記載する.

表 6　「弥報」の評価データにおける Top-1 の具体例

質問	手法	記事タイトル	分類
「コロナ資金繰り支援継続プログラム」とは具体的にどのような支援を提供しているのですか？	FAISS	新型コロナウイルスで売上高減少...	×
	LLM	...「コロナ資金繰り支援継続プログラム」とは	○
	CoT	...「コロナ資金繰り支援継続プログラム」とは	○
業務委託契約と雇用契約の違いは何ですか？具体的な例を挙げて説明してください.	FAISS	... 業務委託契約締結時のポイント	○
	LLM	副業を業務委託でするメリットは？...	×
	CoT	... 業務委託契約締結時のポイント	○
内部留保があると, 会社にとってどのような緊急事態に備えることができるのでしょうか？	FAISS	『日本でいちばん大切にしたい会社』...	×
	LLM	中小企業が美しい決算書を作る意味...	×
	CoT	... 企業の「内部留保」はどのくらいが適切？	○